# Chapter 4

- 4.1 Normal approximation (Laplace's method)
- 4.2 Large-sample theory
- 4.3 Counter examples
- 4.4 Frequency evaluation (not part of the course, but interesting)
- 4.5 Other statistical methods (not part of the course, but interesting)

# Normal approximation (Laplace approximation)

- Often posterior converges to normal distribution when $n \to \infty$
- If posterior is unimodal and close to symmetric
  - we can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

# Normal approximation (Laplace approximation)

- Often posterior converges to normal distribution when $n \to \infty$
- If posterior is unimodal and close to symmetric
  - we can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - ie. log posterior $\log p(\theta|y)$ can be approximated with a quadratic function

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

# Taylor series

- Univariate Taylor series expansion around $x = a$

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots$$

# Taylor series

- Univariate Taylor series expansion around $x = a$

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \dots$$

- Multivariate series expansion

$$f(\mathbf{x}) = f(\mathbf{a}) + \frac{\partial f(\mathbf{x}')}{\partial \mathbf{x}'}_{\mathbf{x}'=\mathbf{a}}(\mathbf{x}-\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})^T \frac{\partial^2 f(\mathbf{x}')}{\partial \mathbf{x}'^2}_{x'=\mathbf{a}}(\mathbf{x}-\mathbf{a}) \dots$$

# Normal approximation

- Taylor series expansion of the log posterior around the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[ \frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \ldots$$

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left( -\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta}) \right)$
- Normal approximation

$$p(\theta|y) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

# Normal approximation

- $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

  - $I(\hat{\theta})$ is the second derivatives at the mode and thus describes the curvature at the mode
  - if the mode is inside the parameter space, $I(\hat{\theta})$ is positive
  - if $\theta$ is a vector, then $I(\theta)$ is a matrix

# Normal approximation – example

- Normal distribution, unknown mean and variance
  - uniform prior $(\mu, \log \sigma)$
  - normal approximation for the posterior of $(\mu, \log \sigma)$

$$\log p(\mu, \log \sigma | y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  - uniform prior $(\mu, \log \sigma)$
  - normal approximation for the posterior of $(\mu, \log \sigma)$

$$\log p(\mu, \log \sigma | y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]$$

first derivatives

$$\frac{d}{d\mu} \log p(\mu, \log \sigma | y) = \frac{n(\bar{y} - \mu)}{\sigma^2},$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  - uniform prior $(\mu, \log \sigma)$
  - normal approximation for the posterior of $(\mu, \log \sigma)$

$$\log p(\mu, \log \sigma | y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]$$

first derivatives

$$\begin{aligned} \frac{d}{d\mu} \log p(\mu, \log \sigma | y) &= \frac{n(\bar{y} - \mu)}{\sigma^2}, \\ \frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}, \end{aligned}$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  - uniform prior $(\mu, \log \sigma)$
  - normal approximation for the posterior of $(\mu, \log \sigma)$

$$\log p(\mu, \log \sigma | y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]$$

first derivatives

$$\begin{aligned} \frac{d}{d\mu} \log p(\mu, \log \sigma | y) &= \frac{n(\bar{y} - \mu)}{\sigma^2}, \\ \frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}, \end{aligned}$$

from which it is easy to compute the mode

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \frac{1}{2} \log \left(\frac{n-1}{n} s^2\right)\right)$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  first derivatives

$$\begin{aligned}
\frac{d}{d\mu} \log p(\mu, \log \sigma | y) &= \frac{n(\bar{y} - \mu)}{\sigma^2}, \\
\frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}
\end{aligned}$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  first derivatives

$$\frac{d}{d\mu}\log p(\mu,\log\sigma|y) = \frac{n(\bar{y}-\mu)}{\sigma^2},$$
$$\frac{d}{d(\log\sigma)}\log p(\mu,\log\sigma|y) = -n+\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{\sigma^2}$$

  second derivatives

$$\frac{d^2}{d\mu^2}\log p(\mu,\log\sigma|y) = -\frac{n}{\sigma^2},$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  first derivatives

$$
\begin{aligned}
\frac{d}{d\mu} \log p(\mu, \log \sigma | y) &= \frac{n(\bar{y} - \mu)}{\sigma^2}, \\
\frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}
\end{aligned}
$$

  second derivatives

$$
\begin{aligned}
\frac{d^2}{d\mu^2} \log p(\mu, \log \sigma | y) &= -\frac{n}{\sigma^2}, \\
\frac{d^2}{d\mu d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -2n \frac{\bar{y} - \mu}{\sigma^2},
\end{aligned}
$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  first derivatives

$$\frac{d}{d\mu}\log p(\mu, \log\sigma|y) = \frac{n(\bar{y}-\mu)}{\sigma^2},$$
$$\frac{d}{d(\log\sigma)}\log p(\mu, \log\sigma|y) = -n + \frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{\sigma^2}$$

  second derivatives

$$\frac{d^2}{d\mu^2}\log p(\mu, \log\sigma|y) = -\frac{n}{\sigma^2},$$
$$\frac{d^2}{d\mu d(\log\sigma)}\log p(\mu, \log\sigma|y) = -2n\frac{\bar{y}-\mu}{\sigma^2},$$
$$\frac{d^2}{d(\log\sigma)^2}\log p(\mu, \log\sigma|y) = -\frac{2}{\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  second derivatives

$$
\begin{aligned}
\frac{d^2}{d\mu^2} \log p(\mu, \log\sigma | y) &= -\frac{n}{\sigma^2}, \\
\frac{d^2}{d\mu(\log\sigma)} \log p(\mu, \log\sigma | y) &= -2n\frac{\bar{y}-\mu}{\sigma^2}, \\
\frac{d^2}{d(\log\sigma)^2} \log p(\mu, \log\sigma | y) &= -\frac{2}{\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)
\end{aligned}
$$

  matrix of the second derivatives at $(\hat{\mu}, \log\hat{\sigma})$

$$
\begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}
$$

# Normal approximation – example

- Normal distribution, unknown mean and variance
  posterior mode

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \frac{1}{2}\log\left(\frac{n-1}{n}s^2\right)\right)$$

matrix of the second derivatives at $(\hat{\mu}, \log \hat{\sigma})$

$$\begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}$$

normal approximation

$$p(\mu, \log \sigma | y) \approx \mathrm{N}\left(\begin{pmatrix} \mu \\ \log \sigma \end{pmatrix} \middle| \begin{pmatrix} \bar{y} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 1/(2n) \end{pmatrix}\right)$$

# Normal approximation – example

- normal approximation can be computed numerically
  - finite-difference for gradients
  - minimize the negative log posterior density: minimum is the mode and Hessian at the minimum is the observed information at the mode
  - e.g. Matlab
    [w,fval,exitflag,output,g,H]=fminunc(@nlogp,w0,opt,x,y,n);

# Bioassay example

| Dose, $x_i$ (log g/ml) | Number of animals, $n_i$ | Number of deaths, $y_i$ |
|---|---|---|
| -0.86 | 5 | 0 |
| -0.30 | 5 | 1 |
| -0.05 | 5 | 3 |
| 0.73 | 5 | 5 |

- $y_i|\theta_i \sim \mathrm{Bin}(n_i, \theta_i)$
- Logistic regression $\mathrm{logit}(\theta_i) = \alpha + \beta x_i$
- Likelihood

  $$p(y_i|\alpha, \beta, n_i, x_i) \propto [\mathrm{logit}^{-1}(\alpha + \beta x_i)]^{y_i}[1 - \mathrm{logit}^{-1}(\alpha + \beta x_i)]^{n_i - y_i}$$

- Posterior

  $$p(\alpha, \beta|y, n, x) \propto p(\alpha, \beta) \prod_{i=1}^{4} p(y_i|\alpha, \beta, n_i, x_i)$$

- demo4_1

# Large sample theory

- Asymptotic normality
  - as $n$ the number of observations $y_i$ increases the posterior converges to normal distribution
  - see counter examples

# Large sample theory

- Assume "true" underlying data distribution $f(y)$
  - observations $y_1, \ldots, y_n$ are independent samples from the joint distribution $f(y)$
  - "true" data distribution $f(y)$ is not always well defined
  - in the following we proceed as if there were true underlying data distribution
  - for the theory the exact form of $f(y)$ is not important as long at it has certain regularity conditions

# Large sample theory

- Consistency
  - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some $\theta_0$, then posterior converges to a point $\theta_0$, when $n \to \infty$

# Large sample theory

- Consistency
  - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some $\theta_0$, then posterior converges to a point $\theta_0$, when $n \to \infty$
- if true distribution is not included in the parametric family, then there is no true $\theta_0$
  - true $\theta_0$ is replcaed with $\theta_0$ which minises the Kullback-Leibler divergence from $f(y)$

$$H(\theta_0) = \int f(y_i) \log \left( \frac{f(y_i)}{p(y_i|\theta_0)} \right) dy_i$$

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Under- and non-identifiability
  - model is under-identifiable, is model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point $\theta_0$ where posterior would converge
  - e.g. if we never observe $u$ and $v$ at the same time and the model is

    $$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

    then correlation $\rho$ is non-identifiable

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Under- and non-identifiability
  - model is under-identifiable, is model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point $\theta_0$ where posterior would converge
  - e.g. if we never observe $u$ and $v$ at the same time and the model is

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

    then correlation $\rho$ is non-identifiable
  - e.g. $u$ and $v$ could be length and weight of a student; if only one of them is measured for each student, then $\rho$ is non-identifiable

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Under- and non-identifiability
  - model is under-identifiable, is model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point $\theta_0$ where posterior would converge
  - e.g. if we never observe $u$ and $v$ at the same time and the model is
  $$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$
  then correlation $\rho$ is non-identifiable
  - e.g. $u$ and $v$ could be length and weight of a student; if only one of them is measured for each student, then $\rho$ is non-identifiable
  - problem also for other inference methods like MCMC

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- If the number of parameter increases as the number of observation increases
  - in some models number of parameters depends on the number of observations
  - e.g. spatial models $y_i \sim \mathrm{N}(\theta_i, \sigma^2)$ and $\theta_i$ has spatial prior
  - posterior of $\theta_i$ does not converge to a point, if additional observations do not bring enough information

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Aliasing (valetoisto)
  - special case of under-identifiability where likelihood repeats in separate points
  - e.g. mixture of normals

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda \,\mathrm{N}(\mu_1, \sigma_1^2) + (1 - \lambda)\,\mathrm{N}(\mu_2, \sigma_2^2)$$

    if $(\mu_1, \mu_2)$ are switched, $(\sigma_1^2, \sigma_2^2)$ are switched and replace $\lambda$ with $(1 - \lambda)$, model is equivalent; posterior would usually have two modes which are mirror images of each other and the posterior does not converge to a single point
  - usually not a big problem for Monte Carlo methods, but may make the convergence diagnostics more difficult

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Unbounded (Rajoittamaton) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume $\lambda$ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any $i$ and $\sigma_1^2 \to 0$, then likelihood $\to \infty$
  - if prior for $\sigma_1^2$ does not go to zero when $\sigma_1^2 \to 0$, then the posterior is unbounded
  - when $n \to \infty$ the number of likelihood modes increases
  - problem for any inference method (e.g. Monte Carlo)
  - can be avoided with good priors
  - note that prior close to a prior allowing unbounded posterior may produce almost unbounded posterior

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Improper posterior
  - asymptotic results assume that probability sums to 1
  - e.g. Binomial model, with Beta$(0, 0)$ prior and observation $y = n$
    - posterior $p(\theta|n, 0) = \theta^{n-1}(1-\theta)^{-1}$
    - when $\theta \to 1$, then $p(\theta|n, 0) \to \infty$
  - problem for any inference method (e.g. Monte Carlo)
  - can be avoided with proper priors
  - note that prior close to a improper prior may produce almost improper posterior

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Prior distribution does not include the convergence point
  - if in discrete case $p(\theta_0) = 0$ or in continuous case $p(\theta) = 0$ in the neighborhood of $\theta_0$, then the convergence results based on the dominance of the likelihood do not hold

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Prior distribution does not include the convergence point
  - if in discrete case $p(\theta_0) = 0$ or in continuous case $p(\theta) = 0$ in the neighborhood of $\theta_0$, then the convergence results based on the dominance of the likelihood do not hold
  - not a problem for Monte Carlo methods (but may still be undesired)
  - should have a positive prior probability/density where needed

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Convergence point at the edge of the parameter space
  - if $\theta_0$ is on the edge of the parameter space, Taylor series expansion has to be truncated, and normal approximation does not necessarily hold
  - e.g. $y_i \sim \mathrm{N}(\theta, 1)$ with a restriction $\theta \geq 0$ and assume that $\theta_0 = 0$
  - posterior of $\theta$ is left truncated normal distribution with $\mu = \bar{y}$
  - in the limit $n \to \infty$ posterior is half normal distribution

# Large sample theory – counter examples

- Does not always hold when $n \to \infty$
- Convergence point at the edge of the parameter space
  - if $\theta_0$ is on the edge of the parameter space, Taylor series expansion has to be truncated, and normal approximation does not necessarily hold
  - e.g. $y_i \sim \mathrm{N}(\theta, 1)$ with a restriction $\theta \geq 0$ and assume that $\theta_0 = 0$
  - posterior of $\theta$ is left truncated normal distribution with $\mu = \bar{y}$
  - in the limit $n \to \infty$ posterior is half normal distribution
  - not a problem for Monte Carlo

# Large sample theory – counter examples

- Tails of the distribution
  - normal approximation may be accurate for the most of the posterior mass, but still be inaccurate for the tails
  - e.g. parameter which is constrained to be positive; given a finite $n$, normal approximation assumes non-zero probability for negative values
- Monte Carlo has different kind of problems with the tails

# Other distributional approximations

- Many other distributional approximations exist and it's a hot research topic in probabilistic machine learning
  - benefit is speed
  - challenge is accuracy and algorithmic robustnes
- Chapter 13 includes
  - more about mode finding
  - more about Laplace approximation
  - variational inference
  - expectation propagation